**LINEAR REGRESSION (INTRODUCTION)**

Decisions in business are often based on predictions of what might happen in the future. Obviously, if a manager’s ability to forecast future events improves, prospects for good decisions improve as well. In this context, “predicting the future” refers to a process that establishes quantitative relations between what is known (*i.e*., that which we can observe, data for example) and what is to be predicted (*i.e*., future outcomes). To that end, we will study regression and correlation; these are interrelated statistical techniques that allow decision makers, not only to estimate quantitative relations among variables, but also measure the “strength” of these relations.

When applying regression analysis, a mathematical equation establishes a theoretical relation among the variables in question; the equation is then estimated statistically. The real question is: “can we establish a pattern in the historical relation between a particular variable of interest (the response variable) and some (explanatory) variable(s) to allow us make reliable predictions of the response variable?” Some very practical examples include estimating: the relation between advertising expenditure and level of demand for a product; the volume of production and material cost; peoples’ smoking habits and the incidents of cancer.

In each of the examples, the relation to be established is of a stochastic[[1]](#footnote-1) nature. This means, for instance, that we do not pretend to imply that the level of demand exclusively and deterministically depends on the *Advertising Budget*. Rather, we simply hypothesize that, *among other factors*, the advertising budget has some non-trivial effect on the level of demand that can mathematically be captured. Consequently, observing the advertising budget does not allow us to know what sales will be without any error; it simply affords us a more accurate prediction than would be possible without that knowledge. It is important to note the difference between establishing a statistical relation and identifying a deterministic relation that is established by scientific laws. That is, while we can only estimate the impact advertising has on demand among other factors, a physicist can tell you very precisely the amount of time it takes a one pound rock to fall 10 feet in a vacuum.

Regression and correlation analyses allow us to establish a quantitative relation or an association between two (or more) variables. The ***independent*** (or explanatory) variable(s) that is (are) assumed to be known is (are) used to predict the ***dependent*** (or response) variable. As this relates to the advertising and demand example, the advertising budget (the variable the company can control) is the independent variable that will be used to predict demand for the product, which is the dependent variable.

It turns out that regression analysis employs only one dependent variable, but can incorporate potentially a large number of independent variables[[2]](#footnote-2). So, in an attempt to better predict sales (the only dependent variable) we might choose to include price in addition to advertising (two independent variables). If only one independent variable is used we refer to the model as a **simple regression model**, whereas **multiple regression models** employ more than one independent variable, sometimes referred to as “right hand side” variables.

**The Simple Regression Model**

A simple regression, which implies a single independent variable, can accommodate most any functional relation between the left-hand side [LHS] (*i.e*., dependent) variable and right-hand side [RHS] (*i.e*., independent) variable. In general, it is easier to think about the relation between the variables as either linear or non-linear (*i.e*., concave, convex or some arbitrary polynomial). In the majority of business applications, as well as other social sciences, a **linear** relation is often assumed.

A *simple* *linear* *regression* equation models the stochastic relation between the dependent and the independent variable as:

|  |  |
| --- | --- |
| Y = A + BX + ∈. | (1.1) |

We will refer to equation (1.1) as the ***true model***; it is “true” because it is assumed that is was obtained from the entire population of potentially observed *Xs* and *Ys* rather than just a sample. Of course, normally in practice our analysis will be restricted to a random sample taken from the population. In Eq. (1.1), *Y* is the dependent variable and *X* is the only independent variable – *Y* and *X* represent the observable data we will be using when we estimate the regression model. Epsilon (*∈*) represents the error term, or the influence of all of the other factors not in our model that impact the dependent variable (*Y*). *A* and *B* are the **intercept** and the **slope** of the regression equation, respectively. Because we will use data (a random set of observations) to *estimate* these parameters, we will also expose ourselves to an additional source of uncertainty (*i.e*., in addition to that which is due to the “other” factors, *∈*) stemming from the sampling errors inherent in estimating these parameters.

 The parameter *A* is the expected value of *Y* when *X =* 0, while the slope parameter, *B,* measures the rate of change in *Y* as *X* changes, or the derivative: *dY/dX.* Generally, we have an idea (*i.e*., a hypothesis) regarding the sign of *B* before we estimate the model. For example, it would seem rational to assume that sales decline as the price of an item increases, but this can only be verified (or refuted) after we estimate the regression. If *B* is positive, we say there is a **direct** relation between the dependent and independent variables, because they move in the same direction; as the independent variable increases (decreases), the dependent variable increases (decreases). Conversely, if the parameter *B* is negative, the relation between the dependent and independent variables is **inverse** – when the independent variable increases the dependent variable decreases and vice versa. Note that we generally will not be able to compute *B* (nor *A* for that matter) precisely since we will estimate it from a random sample. So, after we estimate *B*, it is more than simply looking at the estimated value to judge if the relationship between *Y* and *X* is direct (*B* > 0), inverse (*B* < 0) or if *X* and *Y* are unrelated (*B* = 0). It is entirely possible to obtain, say, a positive estimate from a random sample implying a direct relationship, even if the true value of *B* is zero meaning there is no relationship.

As an example let’s assume *Y* represents a firm’s monthly sales volume and *X* the advertising budget. Because factors other than the advertising budget can affect the firm’s sales volume, a range of *Y* values are possible (*i.e*., the relation is not deterministic) for any given value of *X*. For any particular, fixed value of *X*, the distribution of all *Y* values is referred to as the **conditional distribution** of *Y* and it is closely related to the distribution of the random term, *∈*. The conditional distribution is denoted by *Y|X* and is read: “*Y* given *X”*. For example *Y|X* = $500 refers to the distribution of sales volumes that can occur in all of the months in which the advertising budget was $500.

When applying regression analysis it is generally necessary to make a set of specific assumptions about the conditional distributions of the dependent variable – there is a distribution of *Ys* for each value of *X*. Remember: the dependent variable (*Y*) is what we’re trying to predict.

**Regression Assumptions:**

* **Deterministic *X*.** Values of the independent variable(s) is (are) known

without error.

* **Normality**. All of the conditional distributions are normally distributed. For example, the distribution of sale volumes in all months in which advertising has been, or will ever be, $500, is normal.
* **Homoscedasticity.** All of the conditional distributions, regardless of the value of *X*, have the same variance, *σ*2.
* **Linearity.** The means of the conditional distributions are linearly related to the value of the independent variable. The mean of *Y* for any given value of *X,* (*μY|X* , henceforth) is equal to *A + BX*, implying that the mean of *∈* is zero.
* **Independent errors.** The magnitude of the error, *∈* in one observation has no influence of the magnitude of the error in other observations.

**So, what is *∈* (epsilon)?**

In the true model, *Y = A + BX + ∈*, *Y* is modeled as the sum of a “known” deterministic component, *A + BX*, and a random or stochastic component,*∈.*  While *A + BX* is the mean of the conditional distribution of *Y,* *∈* represents the combined effect of potentially many factors impacting *Y*. The distribution of *Y|X* (*i.e.,* *Y* given a particular value of *X*) and *∈* are closely related because any randomness in *Y* stems solely from *∈* (*A,* B and *X* are fixed*.*) If the assumptions listed above hold, it follows that *∈* is normally distributed with a variance of *σ*2 and a mean of 0. The fact that the mean of *Y|X*, *μY|X* is equal to *A + BX* implies that the expected value of epsilon, *E*(∈) *= 0*. This means that the other factors, the factors unaccounted for in the model, create negative errors that are precisely offset by those creating positive errors. If this were not the case, and the positive and negative errors did not offset, and the mean of *∈* were some non-zero quantity, this fixed quantity would be included in the intercept parameter, *A*,leaving the random component with a zero expectation anyway. Moreover, for the error terms to exhibit true randomness they cannot exhibit any predictability – this alone implies that their mean is zero.

It is the errors that cause the model to be stochastic rather than deterministic. Even if we know *A* and, *B* any specific prediction about *Y* (for some fixed *X*) will nonetheless incorporate some error due to the factors omitted from the model. Of course, it’s reasonable to ask: “Why don’t we include those factors?” Well, it’s likely that we may not be able to identify all of the pertinent factors and, even if we could, some or all of them may not be measurable. For example, it’s clear that a consumer’s opinion about the future can impact his decision regarding the purchase of a particular good at any given time. However, we do not have reliable ways to measure peoples’ expectations as they relate to their consumption.

So, the existence of the other factors embedded in *∈* complicates the task of predicting *Y*. Still, there is yet another complicating factor resulting from the fact that *A* and *B* typically are not known and must be estimated from a random sample of observations. This exposes us to sampling errors – a random sample is unlikely to yield the true values of *A* and B.

To better differentiate between the two sources of errors: 1) resulting from the omitted factors and 2) due to sampling errors, let’s work through an example. The example employs the relation between *Advertising Expenditure* and *Sales Volume* (from above) and we’ll examine this relation within two contrasting scenarios: 1. an ideal world (Utopia), in which we have much more information (*i.e*., data) than we would normally expect and 2. a world much more similar to our own (Reality).

 ***Utopia***: In a utopian world, where everyone is good looking, believes in free markets and we have all of the data in the world, we could make pretty accurate predictions of *Y* given some value of *X*. Suppose we knew (but of course we most likely never will!) the true model relating advertising to sales was *Y* = $649.58 + 1.1422*X*. Now, this means that we have computed the relation between *Advertising* and *Sales* using all of the data that have ever existed or will ever exist – we have all the data – the entire population. Nonetheless, because the relation between advertising and sales is stochastic, any prediction from the model still is likely to be incorrect (when compared with the observed data) because of our ignorance of the other factors that *∈* represents.

Imagine that we want to predict *Sales Volume* when *Advertising Expenditure* is $500. To generate a prediction, we’d simply plug into the model *X* = 500 and calculate *Y* (*Sales*) = $649.58 + 1.1422\*($500) = $1220. This tells us that the mean (*i.e*., expected) value of *Sales* when *X*= $500 is $1220. Or, when $500 is spent on *Advertising*, we can expect, on average, $1220 in sales. So, this prediction is a general statement about the relation between expected *Sales* and *Advertising Expenditure*, and the prediction about the mean will be correct without error.

Suppose, however, that we want to make a specific prediction about *Sales* in a particular month (say next month) if we spend $500 on *Advertising*. It turns out that even with the correct values of *A* and *B,* we cannot make this prediction without error. The error arises because we are trying to identify a particular occurrence of sales as it relates to a specific amount of advertising expense. Sure, on average, we know what happens, but in any particular occurrence all bets are off. In fact, we can feel pretty sure that $1220 is not correct for any one occurrence of sales when advertising expense is $500. Think about it: we know that the average height of an American man is about 5’10”. So, 5’10” is your best guess for a sample of men drawn from the population, but if you randomly pull a guy’s name from a hat it’s very likely he won’t be 5’10” tall.

Predicting the value of a specific occurrence means that we now have to deal with the effect of all *other* factors; we can only say that we expect next month’s sales to be **more or less** $1220, depending on how the other factors (that are embedded in*∈* ) materialize. Of course, the magnitude of “more or less” depends upon the value of *σ*2, which measures the variation (*i.e*., dispersion) of the error around the model’s predictions. Consequently, it is appropriate to think of *σ*2 as a loose measure of the model’s ability to “fit” the data.

In utopia, because we have all of the data, we can compute *σ* without error. In this case, suppose the population data tell us that * =* 356. We can now state that next month’s sales will be within 1220 ± 1.96 \*356 with 95% confidence. Of course, you can state any appropriate confidence interval by selecting the *z* value accordingly.

The point of discussing Utopia, as unlikely a scenario as it is, is to demonstrate the nature of errors and provide some reference for the difference between the two types of errors. By assuming that we have all the data and therefore the true *A* and *B*, we eliminated the sampling errors, but we still have to deal with the errors stemming from the unpredictable impact of all other factors. So, even in the best of worlds, where we have all the data we could ever hope to attain, the model embeds errors, but these errors would only apply to specific predictions about a specific outcome. If we wanted to predict the mean of *Y* for some fixed *X*, we could make this prediction without error, because we know *A* and *B* and *∈*,on average, vanishes. However, making a prediction for a specific occurrence of *Y* for a fixed *X* is still subject to the curse of uncertainty in *∈.*

***Reality****:* In reality, alas, we do not know the true model (*i.e*., *A, B, σ*2). Rather, we have a random sample of observations of *Y* and *X* values from which we can calculate estimates *of A, B, σ*2. We’ll denote estimates of *A, B* and *σ* as *a, b* and *se*.,respectively. Obviously, the fact that we’re using a random sample from the population rather than the entire population makes predicting values of *Y* more difficult, because in addition to the effect of all of the omitted, other factors composing*∈*, we have to consider that our estimates of *A* (using the statistic *a*), *B* (using the statistic *b*) and *σ*(using the statistic se) very likely have errors embedded in them as well. After all, they come from a random sample.

In reality, even predicting the mean of the distribution of *Y* associated with a fixed *X* is no longer guaranteed, unless the sampling errors of (*a – A*) and (*b –B*) happen to be zero, which is very unlikely. [[3]](#footnote-3) Then again, even if the sampling errors were zero, we would have no way of knowing it. As a result, in the more realistic scenario in which we live, even making a prediction about the mean will be hampered by the sampling errors (*a – A*) and (*b –B*); and making a prediction about a specific *Y* will entail these errors plus those associated with all of the omitted factors.

**Using the (Ordinary) Least-squares method of estimating the model**

Suppose instead of having the knowledge of the entire population, we are given a sample set of observations in the form of *n* pairs of *X* and *Y* values; these data are plotted on the following graph. We want to determine the values of *a* (the intercept) and *b* (the slope) so that the sum of the squared vertical differences between the observed *Y* and the estimated *Y* (, where **) is minimized. This process is called the method Ordinary Least-Squares (OLS).



Figure 1

At *Xi* the vertical difference between the estimate, , and the actual observation, *Yi* is

*ei = (Yi -* ). The quantity to be minimized (by choosing *a* and *b*) is, . This is referred to as the sum of squared errors or, *SSE*. It can be proven that if we choose *b* as:

 (1.2)

and *a* as

, (1.3)

then *SSE* will be minimized. Furthermore, the simple sum of the errors, Σ*ei* will vanish (*i.e*., equal zero). Also, an estimate of the *σ*2 is obtained from

. (1.4)

Formulas (1.2), (1.3) and (1.4) give the least-squares estimate of the true model (*i.e*., the true relation between *Y* and *X*) as.

In the formulas for *b* and *se* two equivalent formulas are given: the first one is the so-called “definitional” and the second is the “computational” form, which is more convenient for manual computations. Notice that the OLS line, minimizes *se* by minimizing the numerator [Eq.(1.4)], which is the sum of the squared vertical distances. Also, note how similar *se* is to the formula of a simple standard deviation for a set of observations. When computing the standard deviation for a set of one dimensional (one variable) set of observations, each value is compared to the common mean and the differences are squared. In the least squares formula, each observation is compared to its *own* (conditional) *mean* and the differences are squared.

 To illustrate the method of least squares consider the twelve observations of data (plotted in the above graph) given in the second and third columns of the following table. In order to efficiently compute the statistics *a* and *b*, calculate the required quantities in the computational formulae above as in columns four, five and six.

Table 1: Sample data and some computations

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
|  |  |  |  |  |  |  |  |
| 1 | 573.3 | 1822.4 | 1044764 | 328673 | 3321026 | 1474.0 | 348.4 |
| 2 | 510.6 | 1399.4 | 714523 | 260701 | 1958347 | 1388.2 | 11.2 |
| 3 | 403.9 | 1200.4 | 484826 | 163125 | 1440957 | 1242.2 | -41.8 |
| 4 | 339.6 | 723.1 | 245555 | 115328 | 522830 | 1154.2 | -431.2 |
| 5 | 1289.6 | 2587.2 | 3336357 | 1663034 | 6693352 | 2453.9 | 133.3 |
| 6 | 606.0 | 1495.3 | 906116 | 367200 | 2235964 | 1518.7 | -23.3 |
| 7 | 31.7 | 749.4 | 23753 | 1005 | 561664 | 733.0 | 16.4 |
| 8 | 440.1 | 1516.0 | 667231 | 193720 | 2298142 | 1291.8 | 224.2 |
| 9 | 160.1 | 1197.2 | 191677 | 25632 | 1433378 | 908.7 | 288.6 |
| 10 | 694.4 | 988.7 | 686505 | 482169 | 977436 | 1639.6 | -651.0 |
| 11 | 488.5 | 1357.8 | 663230 | 238588 | 1843660 | 1357.9 | -0.1 |
| 12 | 852.4 | 1981.1 | 1688597 | 726531 | 3924623 | 1855.8 | 125.3 |
| ***Σ*** | 6390 | 17018 | 10653133 | 4565706 | 27211379 | 17017.9 | 0 |
| ***Mean*** | 532.5 | 1418.17 |  |  |  |  |  |

Applying formulas (1.1), (1.2) and (1.3) from above:





.

Notice that the calculations must be performed in this order because *a* requires *b* and *se* requires both *a* and *b*. We also have calculated the predicted *Y* ( [column (7)]) and *ei* [column 8], which sums to zero.

**Sampling Distributions of *a* and *b***

Because these estimates (statistics) are obtained from a random sample, their values are not fixed like *A* and *B* in the true model, but vary from sample to sample; *a* and *b* are themselves random variables. If we had taken a different random sample of 12 observations from the population, we would most likely have computed different estimates for *A, B* and **. To illustrate this concept further, consider two additional samples below that could have been drawn from the same population. Keep in mind that we could have drawn a very large number of such samples, but for our purposes here, we have illustrated only three – the first sample on which we performed the original computations above, plus the two following samples.

Table 2: two additional random samples and their results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X | Y |  | X | Y |
| 538.8 | 891.5 | 705.1 | 1794.7 |
| 520.3 | 922.0 | 259.3 | 1470.9 |
| 267.0 | 211.1 | 132.8 | 1235.6 |
| 572.4 | 556.1 | 462.8 | 775.0 |
| 271.5 | 1463.2 | 1023.0 | 2601.2 |
| 458.3 | 1184.3 | 342.2 | 997.9 |
| 180.4 | 1270.9 | 724.0 | 1642.7 |
| 424.2 | 845.9 | 500.8 | 1448.2 |
| 555.9 | 1299.1 | 295.9 | 952.2 |
| 338.6 | 584.0 | 358.8 | 1368.4 |
| 427.5 | 1119.9 | 770.6 | 1457.2 |
| 884.1 | 1734.3 | 1044.9 | 1800.4 |

  

 *se* = 416.03 *se* = 337.31

 You can imagine the large number of values that we could have computed for *a* and *b*, based on the sample we randomly drew. So, like any other statistic, *a,* *b* and *se*, calculated from a random sample, can be associated with its own probability distribution (*i.e*., frequency distribution of all possible *a* , *b* and *se*  values), called the **sampling distribution**. The standard deviations of the sampling distributions are referred to as the **standard errors** of these estimators and they play a central role in the use of the model for various types of predictions. For instance, if all the possible *b*s happened to cluster around a single value; that is, if the standard deviation of the sampling distribution of *b* (its standard error) was small, we may have more confidence in the specific *b* that we observed in estimating the value of *B*.

Consider an equation relating the distance covered by a train travelling at 100 miles per hour to the duration of a trip. This relation is deterministic in the sense that when we know the amount of time, we can calculate the distance travelled without error. If we plot all possible times against the distance covered, the points will lie perfectly on a straight line, (*i.e*., the fit will be perfect). Therefore, no matter what random sample we examine, all of the data would lie exactly on the same line. All of the samples would lead us to calculate the same least-squares values for *a* and *b*. This means, in the case of a deterministic model, the standard errors of *a* and *b* would both be zero. Whereas when estimating a stochastic relation, we clearly cannot expect that all of the sample observations will lie perfectly on a line and we therefore cannot expect non-zero standard errors. The degree to which the points are scattered around the line (**or **its estimate,*se*  has a strong impact on the standard errors. And that is precisely the reason why the method we used (OLS) minimizes sum of the squared vertical distances or *se*in computing *a* and *b*. We’ll see how these errors are estimated when we discuss making inferences with the estimated model later.

**The Strength of the estimated relationship**

Unlike a deterministic relation, in a stochastic relation the points will not lie perfectly on a line. Rather, the observed *Y* values will be scattered around a line. Of course, this begs a question of fit: if the relation is not perfect, then how good is it? When the relation is stochastic in nature, it’s reasonable to speak of the strength of the relation, or the “goodness of fit.”

Because the degree of fit (strength) is associated with the vertical distances between the observed value of *Y* and the predicted value,, we may consider using *se* [given in Eq. (1.4)] as a measure of strength. But, while this is a reasonable measure, it has an important shortcoming in that it is influenced profoundly by the magnitude of the data – the sheer measurement of the data, whether it is in pennies, dollars or millions of dollars affects *se*. This means that *se* is rendered useless in comparing models from different areas. Regardless of the relative degree of fit, a model in which *Y*s are measured millions will likely result in a much bigger *se*than a model in which the *Y*’s are measured in thousands.

 A better measure, that not only overcomes the influence of the magnitudes of the data but also lends itself to an intuitive and practical interpretation, is the **coefficient of determination**. Simply stated, the coefficient of determination is the proportion of variation in *Y* that can be attributed to the influence of the independent variable, *X*. As a benchmark consider that in a deterministic model in which speed is fixed, all of the variation in the dependent variable (distance travelled) can be attributed to the independent variable (time spent travelling). In other words, the proportion of the variation in distance (*Y*) that can be explained by the variation in time (*X*) is one; all of the variation in *Y* is explained by the variation of *X*. If, conversely, there is no significant relation between *X* and *Y*, all of the variation will be left unexplained leading to a coefficient of determination of zero. For the most part, a stochastic model will produce a coefficient of determination ranging between the two extremes of zero and one.

In regression analysis, the magnitude of variation is measured in terms of the sum of squared deviations (*SS*), which is associated with a “degrees of freedom” (*df*).[[4]](#footnote-4) The overall observed variation in the dependent variable (*Y*) is calculated as the *SST* (Sum of Squares-Total) and it is based on the extent to which each individual value (*Yi*) differs from the mean. Formally, it is represented by equation (1.6), which has *n*-1*df*.

 (1.6)

It is important to recognize that *SST* measures all of the variation in the independent variable, the variable we are trying to predict.

Of course, the idea underlying regression analysis is that there is information in the independent variable that will allow us to predict movements in the dependent variable. Consequently, a reasonable question is: “How much of the variation in the dependent variable (*Y*), *SST*, is attributable to the variation of the independent variable (*X*)?” This is captured by the *SSR* (the Sum of Squares-Regression) and it is given by equation (1.7), which has one *df*.

 (1.7)

So, *SSR* measures the portion of variation in *Y* that canbe explained by the model.

Unless the relation between the dependent and independent variables is deterministic, *SSR* will be less than *SST*, meaning that some of the variation in *Y* remains unexplained and results from other, omitted factors. *SSE* (the Sum of Squares – Error) measures the variation in the dependent variable that remains unexplained; it has *n*-2 *df*.

 (1.8)

For a given set of data, it is always the case that *SST = SSR + SSE*. That is, we partition the variation of *Y* into the part that is explained (*SSR*) and the part that is not (*SSE*). You can see this clearly in Figure 1, above. At the point (*Xi,Yi*), the total “variation” of *Yi* is seen in the vertical distance, , which is clearly positive. This larger-than-average *Sales* volume can partially be attributed to a somewhat larger-than-average *Advertising* budget (*Xi* ) and is given by :the distance between the mean line and regression line. The remaining, unexplained distance, due to other factors, is given by *.*  It is easy to verify that the *df* associated with the three *Sums of Squares (SS)* also are additive: *df*SST = *df*SSR +*df*SSE.

The Coefficient of Determination, *r*2, is simply the ratio of *SSR* (the explained variation in *Y*) to *SST* (the total variation in *Y*), or

 (1.9)

As with the previous expressions, the first formula is definitional while the second is the easier-to-compute by hand “computational” form. Because *r*2 is a proportion, of explained variation to total variation, it will have a value between zero and one, inclusively. A value of one essentially indicates that the relation between the dependent and independent variable is deterministic; all of the variation in *Y* is explained by *X*. If the *r*2 equals zero, there is no relation. Clearly, higher values of *r*2 indicate a better fit of the model to the data.

Let’s employ the data from the previous example to calculate the *r*2 from regressing *Advertising* on *Sales Volume*. Applying equation (1.6) reveals that the *SST* (the total variation in *Y*) is 3,077,335. The *SSE* [Equation (1.8)] shows that 900,722 of the total variation remains unexplained, which implies that 2,176,613 of the variation of *Y* is explained by *X*. The coefficient of determination (*r*2), or the ratio of the explained variation to the total variation, is 2,176,613/3,077,335 = .7073. That is, about 71% of the variability observed in monthly *Sales Volume* is explained by the model (*i.e*., by the variation in *Advertising*).

**Correlation Analysis**

Sometimes it may not be possible to assign specific roles to the two variables—the role of the “response” variable, *Y*, which is assumed to be random, and the role of an “explanatory” variable, *X*, which is assumed to be known with no randomness at all. Suppose we have a random sample of two variables with no basis for assigning these roles, then we must assume they are both samples of random observations (*i.e.,* not just *Y*). In such a situation, we can still speak of the degree of linear association between the two variables in the sense that one of the variable’s changes predictably as the other changes. The variables will be deemed positively correlated if an increase in one of the variables, in most cases, is associated with an increase in the other and *vice versa*; the variables are considered negatively correlated if an increase in one is associated with a decrease in the other. We can measure the strength of the correlation by the **sample correlation coefficient**. Unlike the coefficient of determination, which is always positive, a correlation coefficient ranges from -1 (perfect negative correlation) to +1 (perfect positive correlation).

The correlation coefficient can either be computed as the square-root of the coefficient of determination () appending to it the sign of *b* for obvious reasons; or directly (without running a regression model) as:

 (1.10)

Referencing our previous example, the square root of the coefficient of determination is computed as *r* = √.7073 = +.8410.

 Because *r* is a sample statistic, it estimates the unknown population parameter (*i.e.,* the population correlation coefficient), *ρ*. It is therefore useful to use *r*, computed from the sample, to make inferences about the parameter, *ρ:*

Ho : *ρ =* 0

H1 : *ρ* ≠ 0 ( two-tail)

 **

The appropriate statistic is:

 (1.11)

which has a *student’s* *t* distribution with (*n* -2) *df*. In our example, the expression is computed as

5.

The critical *t* is found under **=.05 with 10 *df*; in this case, 2.2281. Thus, we reject the null hypothesis and conclude that *ρ* is not zero*.* A one-tailed test (to see if either there is a positive correlation or a negative one) uses exactly the same test statistic. However, to concentrate all of the level of significance, ** in the appropriate tail, the critical value is found under 2**one that leaves exactly ** at each tail

**Inferences (predictions) with the Estimated Model**

Once again, the whole point of estimating a regression model is to make better inferences about values of *Y* based on the knowledge of *X*. Before the estimated regression model can be used for this practical purpose however, it’s necessary to verify that there is indeed a statistically significant relation between *Y* and *X*. Lack of a significant relation is tantamount to *B* = 0, which means the true model is *Y = A + ∈*, implying that the knowledge of *X* is completely irrelevant. But, even in cases where there is no statistically significant relation between *X* and *Y* (*B = 0*), it is entirely possible to estimate a non-zero *b* from a random sample because of sampling error*.*  For this reason there must be sufficient evidence in the sample to indicate that *B* is not zero before the estimated model can be used with some confidence.

**Inferences about *B***

Based on the evidence provided by the estimated model, in particular on the value of *b*, we can make two related inferences about *B*: 1) we can create a confidence interval (at some probability level) within which the true *B* lies; and 2) we can test a hypothesis (*i.e.,* an assumption) about the true value of *B* (*e.g*., *B* is zero). Both of these inferences require some knowledge of the sampling distribution of *b,* which can be shown to be normally distributed with a mean of *B* and a standard deviation (*i.e.,* standard error) of

. (1.12)

Intuitively, this means that the statistic, *b*, on average, correctly estimates the parameter *B,* that is *E(b) = B.* When an estimator, on average, correctly estimates a parameter, it is considered an **unbiased estimator** for that parameter*.* A smaller the value of *sb* means that the possible *b* values will be more tightly clustered around *B.*  Of course, we do not have the luxury of knowing the mean of this distribution, but we do know its shape (*i.e*., normal), its standard deviation (*i.e*., the standard error) and we have one observation, the *b* that was computed from the sample.

The statistic (*b - B*) is the error of estimator; it is a random variable because we don’t know *B*. This error can be expressed in terms of the standard error of *b* as

*t*  = (*b - B*)/*sb*. (1.13)

If this ratio is say 2, it corresponds to an error (*i.e.,* the numerator) that is twice the size of the estimator’s standard error, *sb*. If you think about it, values of *b* more than two standard errors from *B* are not very likely to occur.

Notice how similar Eq. (1.13) is to, which, as you should know, has a *student’s t* distribution with *n -* 1 degrees of freedom. Because Eq. (1.13) is the same kind of a thing: it is the distance of an observation from the mean of a normal distribution expressed in terms of the standard deviation; it also has a *t* distribution, but in this case with *n* – 2 degrees[[5]](#footnote-5) of freedom:

In general, we can calculate a (1*- *) confidence interval for *B* as

 *b* ± *t**sb*. (1.14)

In expression (1.14) *t* is the corresponding value of the *t* distribution with *n -* 2 degrees of freedom for tail probabilities of **. To calculate a 95% confidence interval for *B* based on the observed statistic *b* = 1.368, it is necessary to compute the standard error of *b* from equation (1.12) as .2783. The value of *t*.05 with 10 *df* (*i.e*.,

*n* = 12 and we are estimating 2 parameters, *a* and *b*) is 2.2281 from the *t* table. Accordingly, the confidence interval for *B* is 1.368 ± 2.2281\* .2783. This means that, given the sample, there is a 95% probability that the true *B* lies in the interval 0.7480 to 1.9881. It is significant that the interval does not include zero, which provides some confidence that the true *B* is not zero and that *X* has a significant impact on *Y*.

 The second type of inference about *B* judges if its estimated value, *b,* contradicts an a priori value. Most commonly, this is used to conclude that *B* is statistically different from zero by attempting to contradict the null hypothesis that *B* is equal to zero. If we can reject the null hypothesis at some level of significance, denoted by ** we can then use the estimated model to make predictions of *Y* for a given *X.*

As usual, we form two competing hypotheses. The null hypothesis, Ho, asserts that *B =* 0, however the alternative hypothesis, H1, may take one of three different forms. If there is a priori reason to believe that *B* cannot possibly be negative, then H1 asserts *B* > 0. Conversely, if there is a priori reason to rule out that *B* is positive, then H1: *B* < 0. Finally, when there is no basis for ruling out either tail, then H1: *B* ≠ 0.

To test the validity of the competing hypotheses, Ho and H1,apply equation (1.13) to compute the test statistic and then establish if the statistic is too extreme to be consistent with the null hypothesis.

**Three possible test scenarios:**

1. Ho : *B =* 0

 H1 : *B* ≠ 0 ( two-tail)

 **

The relevant test statistic is *(b - B*)/*sb* = (1.368 - 0)/0.2783 = 4.9156 and has a *t* distribution (with *n* - 2 *df*), if the null is true. Since we are trying to determine if *B <* 0 *or* *B >* 0, it is necessary to compare 4.9156 to the *t* value with (*n –* 2) *df* in the *t* table. The table splits the level of significance, ****equally between the two tails, so there’s 2.5% in each tail. In this case, the value we want is found under **=.05 with 10 *df*; it’s 2.2281. This means that if the null were true, a *t* value as extreme as ± 2.2281 is no more probable than .05. Because 4.9156 is more extreme than the critical value in the *t* table, we reject the null hypothesis. This is evidence that there is a statistically significant relation between *X* and *Y*. Notice that this is precisely the same test for *ρ* = 0. Concluding that *B* is not zero is the same phenomenon as concluding that there is a statistically significant correlation between the variables.

Ho : *B =* 0

H1 : *B* > 0 (one tail)

**

Once again the appropriate test statistic is [equation (1.13)] *t = b/sb* = 4.9156. Here, we want to determine if the computed *t* is both positive and extreme enough to reject the null hypothesis. Obviously, if the estimated *b* were negative, the alternative could be immediately rejected, because a negative *b* cannot provide evidence for *B >* 0. Now, however, the appropriate *t* value places all of the ** in the right tail.

Most *t* tables are constructed with a two-tailed test in mind; consequently, half of the ** is located in each tail. When a column in the table is designated **.05, the column identifies *t* values consistent with 2.5% in the tails. To accommodate this convention, it is necessary to find the appropriate *t* value under the column 2**; the negative half of the distribution is irrelevant.

In this case, the relevant *t* value is found under **.10, with (*n* - 2) = 10 *df*; it’s value is 1.8125. Once again the null is rejected because the computed *t* statistic of 4.9156 exceeds the table’s critical *t* value of 1.8125.

3. Ho : *B =* 0

H1 : *B* < 0 (one tail)

**

To test the null hypothesis in this case, simply examine the other tail of the *t* distribution. If the *t* value (4.9156) were smaller than -1.8125, then there would be justification to reject the null. In this the example, the estimated *b* of 1.368 allows us to reject the alternative out of hand because the positive *b* value clearly cannot provide evidence negative *and* statistically significant.

**Tests regarding a non-zero *B***

 Sometimes it’s necessary to test whether there is evidence in the sample regarding a specific value of *B*. For example, the Capital Asset Pricing Model (CAPM) makes a specific statement about the risk of an asset and its relation to the market. The CAPM holds that the market’s beta is equal to one; any asset with *B* > 1 is riskier than the market, a *B* < 1 indicates that the asset is less risky. So, when estimating a stock’s beta, it is not uncommon to determine if there is evidence to indicate that is, in fact, statistically different in risk from the market, *B* = 1.

 Regardless, nothing much changes: it remains appropriate to apply the test statistic (1.13) and the testing method is the same. However, the null hypothesis now asserts that the particular value of *B* = 1.

Ho : *B =* 1

H1 : *B* > 1 (one tail)

**

In this case, *t* = (*b –* *B*)/*sb* = (1.368 – 1)/0.2783 = 1.3223. Because, as we determined earlier, the appropriate *t* for a one tail test at ** = .05 with 10 *df*is 1.8125, we cannot reject the null hypothesis. Although there is some evidence in the sample that *B* > 1, because *b* = 1.368, the evidence is not strong enough to assert *B* > 1 with some confidence.

**What about the intercept, A?**

 We can perform the same tests on the true intercept, *A*, using the exact same approach, which based on its sampling distributions. In most cases, however, this is not as pertinent as inferences about *B*. Therefore, we won’t provide a discussion of inference tests regarding *A*.

**Inferences about *Y*,given some *X = Xo***

Once a statistically significant relation between *X* and *Y* has been established, two kinds of predictions can be made about *Y*: 1. *μY|X* (the mean of *Y|X*) and 2. about *Y|X* ( a specific occurrence of *Y*). To visualize an inference regarding the mean, *μY|X*, imagine the *Advertising* budget is set at $500. Even though the *Advertising* budget remains $500, *Sales* will fluctuate because of other factors. When predicting the mean, we’re simply trying to generally determine what happens on average when the *Advertising* budget is some given value. The question is: “What will our average *Sales* be when *Advertising* is $500?”Predictions about *Y|X* are more challenging. In this case, we’re attempting to make a very specific prediction about a single occurrence. For example, consider predicting next month’s *Sales* volume when the *Advertising* budget will be $500. The question becomes: “What will our *Sales* be next month if we spend $500 on *Advertising*?”

 Both types of inference will be subject to sampling errors. When making inferences about the conditional mean, we are essentially estimating *μY|X*, = *A + BX*, because the mean of *∈* is zero – all of the other factors will, on average, vanish. Any error results only from the fact that we have to estimate *A* and *B* with *a* and *b*, so in essence, the error results from (*a – A*) and (*b – B*). When predicting specific outcomes (*Y|X*), however, the task becomes more cumbersome because the errors associated with the other factors stick around. In essence, the prediction now involves estimating *Y|X = A + BX +∈*. That is, in this case the error results from the sampling error in estimating *A* and *B* **plus** the random play of the omitted factors captured by*∈*. Notice that the point estimate for both *μY|X* and *Y|X* is , but, as a result of the impact of *∈*, the standard error of the statistic will be larger for *Y|X* than for *μY|X*. So, the only difference in the inferences lies in the standard error!

**Confidence interval for *μY|X***

Suppose someone asks: “If the *Advertising* budget is fixed at some *Xo*, what is our mean or expected *Sales* volume?” Obviously, this is a question about what generally happens to *Sales* when *X0* is spent on *Advertising*. As usual, the best, unbiased predictor of what is expected to happen is the average. So, the point estimate is . The standard error of in estimating ***μY|X*** is given by

 . (1.15)

A (1 – α) confidence interval for ***μY|X*** is given by:

  (1.16)

where, as before, *tα* is the student’s *t* value with *n -* 2 degrees of freedom.

As an example, let’s calculate a 90% confidence interval (α = .10) for ***μY|X*** when *X* = $500. The point estimate is (from the regression output, above). The relevant standard error is: **

The t value for α =.10 with 10 degrees of freedom is 1.8125 (5% in each tail). We can now form the 90% confidence interval for ***μY|X* =500** as 1373.69 ± 1.8125\*87.1083, which means that *Sales* can be expected to be somewhere between $1215.81 and $1561.54 - 90% of the time.

**Confidence interval for *Y|X***

Suppose someone asks the question: “If the *Advertising* budget is fixed at some *Xo* in July, what will July’s *Sales Volume* be?” Once again, the best, unbiased estimate of what will happen is the average. So, the point estimate would again be. However, because the question relates to a specific outcome, *Sales* in July, the standard error of in estimating *Y|X* is now given by

. (1.16)

A (1 – α) confidence interval for *Y|X* is given by: . Notice that *sp* is larger than *sµ* and differs by the addition of a 1 under the radicand. This, loosely speaking, adds to the standard error a quantity equal to *se*which, as you by now know, is a measure of the uncertainty created by the existence of all other factors. To compute a 90% confidence interval for *Y*|*X*=$500, we now need **  The confidence interval for *Y*|*X*=$500 is 1373.69 ± 1.8125\* 312.5061 or, when if $500 is spent on *Advertising*, we can expect *Sales* from $807.28 to $1940.09 in July 90% of the time.

 The standard error of both types of inferences include the term , which means that predictions of either type get less accurate as *Xo* deviates more from the mean of the *X* values in the sample. So, the confidence intervals become wider as *X0* is moved further to either side of . This is illustrated in the following figure.



 Figure 2

Think of the gray area in Figure 2 as a 95% confidence interval for *μY|X* given various levels of X, while the adjacent white area represents a 95% confidence interval for *Y|X*.

**Regression Caveats**

It is necessary to use caution when using regression analysis to make inferences. Here are some important points:

1. It is a good practice to make sure that the regression assumptions are not grossly violated; otherwise various inferences made from the model will be invalid, or even misleading.
2. Never infer causality based on a regression. If a regression is estimated and a significant relation between *Y* and *X* is found, it is not appropriate to conclude, based on the model alone, that changes in *X* cause *Y* to change.

For example, if you found a “good fit” between consumers’ spending on high priced jewelry and expensive vacations, it would be incorrect to conclude that people who buy expensive jewelry take expensive trips to show them off. In reality, buying jewelry may have no impact on vacation plans. A more plausible explanation might be related to a third factor, say wealth, which induces spending on expensive jewelry and taking luxury vacations. Here the causal factor is wealth and both jewelry purchases and vacations are being “caused” by it.

1. It is not appropriate to make predictions about *Y* (the mean or a specific value) based on values of *X* that lie outside the range of *X* values in the sample.
1. A stochastic process involves a randomly determined sequence of observations each of which is considered as a sample of one element from a probability distribution. That is, it has an element of randomness that cannot be completely predicted. [↑](#footnote-ref-1)
2. The number of independent variables technically is limited by the size of your sample (at most, you can only use *n*-1 independent variables in a regression) as well as the “rule of parsimony,” which states that a model should be made as simple as possible (*i.e*., use the fewest possible number of variables). [↑](#footnote-ref-2)
3. A sampling error is the error caused by using a sample rather than the entire population. [↑](#footnote-ref-3)
4. Statisticians use the term "degrees of freedom" to describe the number of values in the final calculation of a statistic that are free to vary. Consider, for example the sample variance. To calculate the variance of a random sample, the mean of that sample must be calculated, then the sum of the squared deviations from that mean. While there will be *n* such squared deviations, only (*n* - 1) of them are, in fact, free to assume any value whatsoever. This is because the final squared deviation from the mean must include the *one value of X that allows the sum of all the Xs divided by n to be equal to the obtained mean of the sample*. From *"Glossary of Statistical Terms"*. <http://www.animatedsoftware.com/statglos/sgdegree.htm>. Retrieved on 8/15/2009. [↑](#footnote-ref-4)
5. Two degrees of freedom are lost because the two statistics, *a* and *b*, are computed from the sample. [↑](#footnote-ref-5)